How Automated Agents Treat Humans and Other Automated Agents in Situations of Inequity: An Experimental Study

(Short Paper)

Ron Katz, Sarit Kraus
Bar-Ilan University, Ramat-Gan 52900, Israel, sarit@cs.biu.ac.il

ABSTRACT

This paper explores the question of how agent designers perceive and treat their agent's opponents. In particular, it examines the influence of the opponent's identity (human vs. automated agent) in negotiations. We empirically demonstrate that when people interact spontaneously they treat human opponents differently than automated agents in the context of equity and fairness considerations. However, these difference vanish when people design and implement agents that will interact on their behalf. Nevertheless, the commitment of the agents to honor agreements with people is higher than their commitment to other agents. In the experiments, which comprised 147 computer science students, we used the Colored Trails game as the negotiation environment. We suggest possible explanations for the relationships among online players, agent designers, human opponents and automated opponents.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Multiagent systems

General Terms

Human Factors

Keywords

Human-Agent Interaction, Automated Negotiators

1. INTRODUCTION

The development of efficient and beneficial automated negotiating agents has received an increasing amount of attention, both in academic research (e.g. [3]) and in E-commerce markets, such as eBay (see: pages.ebay.com/help/buy/proxy-bidding.html).Since the agents we design are designated for negotiating with other agents designed by other people, when designing efficient agents we must consider how other people design their agents. Most papers that discuss the question of how to develop efficient agents try to learn and model their opponents in one specific negotiation domain during the interaction (e.g. [2]). However, not many have examined more general questions of how people design their agents to perform on their behalf, and which considerations direct them in the design process [7]. Moreover, as far as we know, no previous research has checked how agent designers perceive their agents' opponents, and how their perception affects the agent's design, compared to direct opponents. Are they more hostile or maybe more

compliant; do they cheat more or are they more patient with opponents that face their agents? Understanding these questions can help us design agents that will cope efficiently with their opponents in real-world markets, especially in markets where most agents are unprofessional and are designed by private users.

In order to explore the perception of agents' opponents by agent designers we decided to focus on one phenomenon that has already been shown to be significant in human-agent interaction literature. Empirical evidence shows that when people interact or play online, they may treat an automated agent differently from a human opponent, in the very same domain. A previous experiment using the Ultimatum Game (UG) [5] revealed a significant difference between the reaction of people to automated opponents and to human opponents [9]. In the UG, two players are given the opportunity to split a sum of money. One player, the proposer, has to make an offer as to how this money should be split between the two. The second player, the responder, can either accept or reject this offer. If it is accepted, the money is split as proposed, but if the responder rejects the offer, then neither player receives anything. In the experiment we have shown that inequitable offers of $2 and $1 (out of $10) made by human proposers are rejected at a significantly higher rate than those offers made by a computerized agent. This finding indicates that participants perceive human and computerized opponents differently, and have a stronger emotional reaction to inequitable offers made by humans than to the same offers from a computer. In another research, similar differences were found between human proposers and a computerized roulette-wheel [1].

In this paper we examine the question of how "agents" perceive their opponents, i.e. how people treat human opponents and computerized opponents when they design and implement an agent for negotiating on their behalf. In particular, we check whether there is a difference between agents designated to interact with human opponents and agents designated to interact with other agents. In order to examine this question we asked two groups of subjects to design and implement agents for interacting either with people or with other agents, in the same negotiation environment. Theoretically, we could have asked the participants to design agents which would play the responder in the UG, similar to the procedure used by Sanfey et al. [9]. However, this would have been a rudimentary task of merely setting an acceptance threshold, and would not have been inherently different from the decision making process made directly by people. Therefore, we have designed a more complex negotiation environment which preserves the basic conflict of a responder in the UG: emotions of equity and fairness on the one hand, and strives to gain some money on the other hand. Our environment is based on the Colored Trails (CT) testbed, which will be broadly introduced later in this paper.

In our experiments we first show that when people interact online (not as agent designers) in our CT environment, they treat human opponents differently than automated agents in the context of equity and fairness considerations. Consistent with Sanfey et al. [9], in our experiments the participants were more willing to cooperate

with an inconsiderate opponent when it was presented as an automated agent, than when it was a human opponent. This difference, however, vanished when we asked people to design and implement agents that would interact on their behalf. In this case, the cooperation with the inconsiderate opponent was high both with a human and with an automated opponent. Namely, when designing agents people prefer to gain more money, and pay less attention to the inconsiderateness of the other player, no matter whether he is human or an automated agent. Nevertheless, the commitment of the agent to honor agreements with people was higher than their commitment to other agents. In addition, we found that when describing the game the agent designers refer to automated opponents using more competitive expressions (such as "opponent" and "adversary") than they refer to human opponents.

The importance of the questions raised in this paper is twofold: First, it is necessary to understand the psychological and social effects of the computer environment on human users, since it has become a dominant framework where a great deal of our social relations and decisions are taking place. Second, understanding these behavioral issues, which are usually overlooked by automated agent designers, can contribute to the design of agents that interact with agents designed by other people.

The concept of the *strategy method* [8] appears in literature on economics. The strategy method is an experimental procedure for eliciting a complete strategy of play for all information sets, not only the ones that happen to be reached during the course of a play of a game. Many papers examine strategy methods of the decision makers and compare them to their spontaneous decision making in the same situation. In the UG, for example, a meta-analysis of 37 papers found that when the strategy method is employed, proposers are more generous and responders are less willing to accept an offer than in spontaneous games (where the responders make their decisions after receiving the offer) [8]. Principally, we believe that the strategy method procedure is quite different from computerized agent programming, since programming grants the ability and the inspiration to use a large number of memory units and sophisticated tools, such as statistical computation and machine learning methods. It is usually very difficult to handle long and multi-optional negotiation protocols using the strategy method procedure. Moreover, it is possible that from the psychological aspect agent designers perceive their agents as autonomous negotiators, and allow them to use different attitudes towards their opponents.

In the next section, we describe the CT negotiation environment that we used for running our experiments. In the subsequent section we present the experimental setting. We then present the results and discuss our findings. In the last section we briefly conclude and outline directions for future work.

2. THE NEGOTIATION ENVIRONMENT

We used the Colored Trails (CT) game [4] as the negotiation paradigm in our experiments. CT is a conceptually simple but highly expressive computer mediated game framework that can be used to model a range of multi player task settings and decision-making situations. The game is played on an NxM (4x4 in the current research) board of different colored squares, as shown in **Fig. 1**. Players move from a starting position toward a goal square using chips of colors that match the board squares. For a player to move into an adjacent square, she must turn in a chip of the same color as the square. Specifically in the current paper, as depicted in Fig. 1 which presents the initial setting of the board, the "me" player can make it to the goal square with the initial allocation of chips and is thus "independent". The other player, represented by the yellow sun icon, needs to obtain the pink and the orange chips from the other player in order to reach the goal.

The chips may be exchanged between the players, according to the communication and negotiation strategies of the game protocol. Each player can send offers for the exchange of chips to the other player, during certain negotiating periods, which can be accepted,
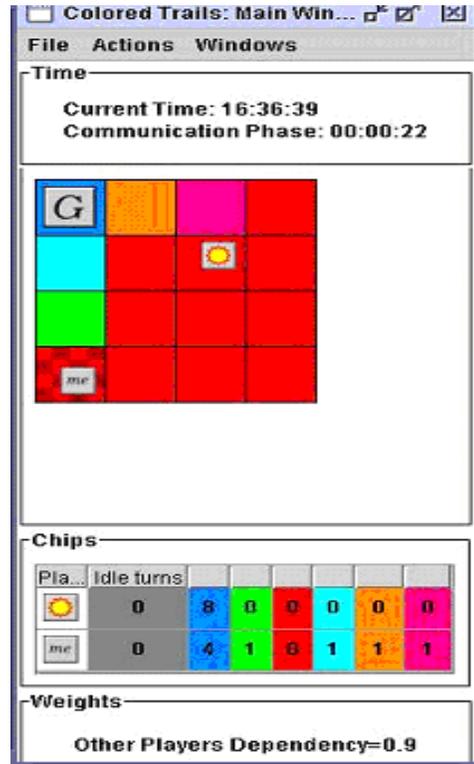

Figure 1: The initial setting of the CT board. The upper section indicates the time that has passed from the beginning of the game. The section below presents the main game board: the location of each of the 2 players, and the goal cell ("G"). The third section describes the chips' inventory of each of the players. Each entry indicates the number of chips of each color. The upper player ("sun") possesses 8 blue chips and the lower one ("Me") representing the examinee, possesses 4 blue, 1 green, 1 azure, 1 orange and 1 pink chip. The other player must essentially acquire the two latter chips, i.e. 1 orange and 1 pink chip, in order to reach the goal. The lower section indicates the portion each player receives from his opponent's final score, which is 0.9.

rejected or unanswered by the receiver of the offer. However, agreements are not enforced by the game controller, allowing players to break agreements (as in "real-life" domains).

A player's performance in CT is determined by the scoring function, which is also set by the game protocol for a particular instantiation of CT. In the current paper the scoring rule gave 200 points to a player who reached the goal square, 15 points for every chip the player possessed at the end of the game, and a deduction of 8 points for every square in the Manhattan distance of the player from the goal square, in case the goal was not reached. In addition to this basic score, each player received the score of the other player multiplied by 0.9. This addition gave the independent player a significant incentive to help the other player reach the goal, by supplying the critical chips needed by the other player. Moreover, it was clarified that performance of individuals would be measured non-competitively; players were to try to maximize their own scores and not minimize other players' scores. Both players had knowledge of the scoring function, as well as a full view of the board and the other player's chips.

In order to place our participants in a similar situation to that of a UG responder facing an inequitable proposal, we paired them with a *non-reciprocal* automated agent. Specifically, we designed and implemented a greedy agent that played the role of the dependent player. This non-reciprocal agent agreed only to accept chips but not to send chips, though it had 7 spare red chips that it could send in return for the 2 chips it needed. Playing as the independent

player against this agent placed the player in a confusing dilemma: on the one hand, it is very profitable to send the 2 chips and thus earn 90% of the 200 points that his opponent (the dependent) would receive for reaching the goal. On the other hand, it is inequitable that the independent player help his opponent to reach his goal, while the opponent exploits the situation and gives him nothing in return. Moreover, although the independent player began the game in a much better situation than the dependent player, if he would send the latter the chips it needs - its final score would be higher than that of the independent player. This conflict between emotions of equity and rational calculation of utility, is quite similar to the conflict of the responder in the UG between rejecting an inequitable proposal and accepting any positive amount of money.

Using the CT environment enabled us to examine how people deal with the conflict between emotions and rationality that underlies the UG, when they design an automated agent. In addition, we were able to validate the findings of [9], concerning the different treatment of humans and of computerized opponents, in a more complex and natural domain. The advantages of the CT framework, which is more similar to real-life negotiations than typical economic games, and can provide an analogue for more complex task settings have, in a very short time, made it a paradigm for various behavioral aspects, such as negotiation orientations [6] and social dependencies [4].

3. EXPERIMENTAL SETTING

Our experiments consisted of 147 subjects, who were divided into four groups. All the groups played or designed agents to play the same CT game as specified in the previous section. However, each of the four groups participated in a unique setting, as described below:
1. Subjects played online against non-reciprocal automated agents, and were aware of the fact that their opponent was an automated agent.
2. Subjects played online against non-reciprocal automated agents, and were led to believe they were playing against another person.
3. Subjects were requested to design and implement an agent that would play on their behalf against another agent. Eventually, we paired their agents with the non-reciprocal agent.
4. Subjects were requested to design and implement an agent that would play on their behalf against a human opponent. Eventually, we paired their agents with the non-reciprocal agent.

The first two groups, which included 36 participants each, served as the control groups. Since we assume that subjects who face the CT scenario experience a similar conflict as the responders in the UG in [9], we expected similar results to those of [9]; i.e., subjects in group 1 playing against automated agents should be more generous toward their opponents than subjects in group 2 who believe they are playing against other people. Thus, more people in the first group should send the two chips that will allow their opponents to reach the goal.

The experiments of groups 1 and 2 consisted of meetings with 8 or 12 subjects each. The subjects were seated apart at computer-terminal stations in a large computer laboratory, in a setting which did not allow them to see each others' screens. The experiment began with a 50 minute oral tutorial of the game accompanied by a written manual, consisting of an explanation of the rules and the scoring function. In addition, all the participants participated in a short and simple practice game of a totally different setting of the CT testbed, which did not include any emotional-rational conflict. At the end of this phase, the experimenters, who were not aware of the goals of the experiment, verified the understanding of the game on the part of each of the participants personally. Each game began with 3 minutes dedicated to becoming familiar with the board setting and contemplating optional tactics. After playing the game, the subjects were asked to fill out a short demographic questionnaire and to briefly summarize the tactics they used in the game.
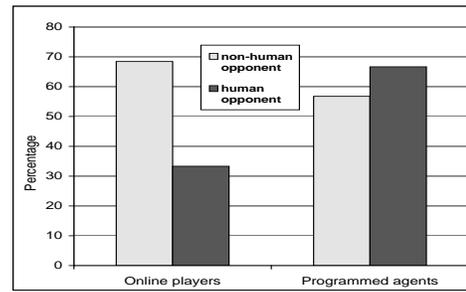

Figure 2: The percentage of players who sent the pink and orange chips to their opponents. The left side presents the results for players who played online, and the right side refers to the programmed agents. The dark bars represent results for players who thought they were playing against human opponents, while the bright bars represent players that were informed they were playing with automated agents.

Groups 3 and 4 included 37 and 38 participants, respectively. Each participant was sent a detailed written manual to his e-mail box, which consisted of an explanation of the game rules and the scoring function. The document for group 3 explained that the designed agents would play against another agent, and for group 4 it stated that they would play against human opponents. The participants also received the code of the game environment written in Java, as well as a detailed document explaining the structure of the system: the classes and the interfaces of their functions. The participants were required to design their agent's strategy and implement the code of the "myAgent" class in accordance. We provided them with a skeleton of the "myAgent" class which included the format of interfaces of the members' methods, as well as definitions of the members' variables. These methods control the sending of chips and offers of exchange to the opponent and the response to the offers coming from the opponent. In addition, the participants were required to submit a verbal description of their agents' strategies.

The participants were all upperclass computer science students at Bar Ilan University, who were not experts in negotiation strategies nor in economic theories directly relevant to the game (e.g., game theory, decision theory). Participation was part of the requirements of a course, and the students were given their grades according to their (or their agents') performance in the game.

4. RESULTS

Examining the left side of **Fig. 2** reveals that the online participants were much more cooperative when they knew they were playing against an automated agent, i.e. 66.67% of the players sent the 2 chips for free, knowing that their opponent was non-human, vs. only 33.33% of the players who thought they were playing against a human opponent (χ^2 test, p $<$.01). This finding is fully compatible with the results of [9] in UG. The designed agents, on the other hand, showed no significant difference in their attitude towards human and non-human opponents, as depicted on the right side of Fig. 2. 56.76% of the agents that played against other agents (group 3) and 68.42% of the agents that played against people (group 4), sent the 2 chips for free.

However, when we examined the agents' code we found a significant gap in the reliability of the agents, i.e. the commitment of the agents to honor agreements reached with the other player. While about 24% of the agents from group 3 (playing against agents) were not fully reliable, and they would deviate from certain agreements they had reached, only 5% of group 4 (playing against people) would not always send all the chips they had agreed upon (χ^2 test, p $<$.05).[1] Another interesting difference was an expressional

---
[1]Examining the reliability of the online players (groups 1,2) was not possible, since they played against the non-reciprocal agent that did not confirm any agreement.

difference found in the literal descriptions of the agents' strategies. When examining the denominations that the participants used in their summaries, when they referred to their opponents, 62% of group 3 vs. only 39% of group 4 used neutral phrases such as: "the other player", "he", "the other agent" and even "my partner". On the other hand, the other 38% of group 3 vs. 61% of group 4, used negative and competitive phrases, such as: "the opponent" and "my adversary" (χ^2 test, p <.05). Recall that in the game description that we sent to the students, we referred to their opponents only as "the other player". [2]

We were also interested in analyzing the results from the perspective of the differences between online players and designed agents. Comparing the rate of agents who sent chips for free from group 1 (online players) with agents from group 3 (designed agents) and agents from group 2 (online players) with agents from group 4 (designed agents) reveals that while online players and designed agents demonstrated no significant difference when playing against agents (groups 1,3), a statistically significant difference was found among the agents playing against people (groups 2,4 - χ^2 test, p <.01). Online players were much less cooperative with non-reciprocal human opponents than with designed agents.

5. DISCUSSION

In consistence with our prediction, the online players in the CT game demonstrated the same pattern of behavior as the UG responders in [9]. A possible explanation can attribute the lower rate of chips sent to the human players' ego or to equity considerations, which were much less for the non-human opponents. In the agents' design experiment (group 3,4), however, this difference totally vanished. One explanation can attribute this phenomenon to the fact that when the participants were in their homes with plenty of time to think and design their agents, they made their decisions more rationally in the sense of maximizing their final score in the game. For this reason more people decided to send chips for free, no matter who their opponent. Another explanation could be the mediation of the agent in the negotiation process. When the participants played online against other people, they may have felt anger or unfairness in reference to their non-cooperative opponents which caused many (66%) to punish their opponents (which caused damage to themselves, as well). However, when agents "handle" the negotiation process on behalf of them, the designers may be less emotionally involved and less sensitive to the opponents' behavior, whether they be human-beings or not.

Nevertheless, when it came to aspects of reliability, truthfulness and even an implicit perception of the opponent as reflected in the verbal summaries of agent designers - there was significant importance to the identity of the opponent. Apparently, despite the agent's mediation, people do feel empathy toward their opponents, when they know they are human-beings.

Another finding that can be inferred from the experiments is the fact that online players are less likely to send chips to human opponents than to programmed agents. This finding may seem inconsistent with [8] who found that responders are less willing to accept an offer when the strategy method is employed. In the same manner, it seems in contradiction with Guth and Tietz [5] who claim that in the UG, using strategy methods strengthens fairness considerations. A possible explanation can ascribe this inconsistency to the difference between strategy methods and agent design. As mentioned in the introduction, since the agent plays on their behalf, less ego considerations are involved among the designers, and their strategy may be less equitable but more utility-maximized. When stating a strategy method, on the other hand, people may want to declare

---

[2] When checking the denominations written by the online players, we found that only two players from group 2 and not a single player from group 1 used negative or competitive phrases. This perhaps could be explained by the fact that the online players were exposed to oral guidance of the experimenters who used the phrase "the other player" over and over again.

a proud and non-servile position, even more than an online player that can change his position during the game. Another explanation is related to the difference between the paradigm of the UG that was used in [8, 5] and the CT game that was examined in our work. In the UG, an extremely unequal division of the money may seem very inequitable, because the money is designated for both of the players. In our CT setting, however, sending the chips for free may be perceived as less inequitable, since it is part of the bargaining conducted between two opponents.

6. SUMMARY

In this paper we explored one aspect of the way agent designers perceive their opponents, i.e. how the opponents' identity influences the design of the agents. This was done by experimentally examining the existence of differences in the way programmed agents treat other agents and human opponents in negotiations. Our findings show that unlike online players who are less sensitive to uncooperative human opponents, the programmed agents show no difference toward agents and human uncooperative opponents. In general, however, agent designers are more reliable and perceive human opponents more positively than artificial opponents.

In the current study we have focused on negotiating with uncooperative opponents, which involves common characteristics of negotiations, such as: cooperativeness, fairness, equity and concern about the outcome. In the future, we intend to examine agents' and human opponents' treatment of other aspects of negotiations, such as: competitiveness, altruism, aggressiveness and reciprocity. Understanding these aspects better can be beneficial when interacting online or when designing automated agents. In addition, it would be interesting to examine how continuous interaction with automated agents influences treatment towards agents, both of online players and of agent designers. We believe there is a great possibility that when people become accustomed to interacting with automated agents, they may personify them and treat them in the same manner they treat human opponents. This assumption, if experimentally demonstrated, may significantly contribute to our understanding of the human-agent interaction world.

7. REFERENCES

[1] S. Blount. When social outcomes aren't fair: The effect of casual attributions on preferences. *Organizational Behavior and Human Decision Processes*, 63:131–144, 1995.

[2] D. Carmel and S. Markovitch. Model-based learning of interaction strategies in multi-agent systems. *JETAI*, 10(3):309–332, 1998.

[3] M. Fasli and O. Shehory. *Agent-Mediated Electronic Commerce. Automated Negotiation and Strategy Design for Electronic Markets*. Springer, 2007.

[4] B. Grosz, S. Kraus, S. Talman, B. Stossel, and M. Havlin. The influence of social dependencies on decision-making: Initial investigations with a new game. In *Proc. of AAMAS'04*, pages 780–787, 2004.

[5] W. Guth and R. Tietz. Ultimatum bargaining behavior: a survey and comparison. *Journal of Economic Psychology*, 11(3):417–449, 1990.

[6] R. Katz, Y. Amichai-Hamburger, E. Manisterski, and S. Kraus. Different orientations of males and females in computer-mediated negotiation. *Computers in Human Behavior*, 24(2):516–534, 2008.

[7] E. Manisterski, R. Katz, and S. Kraus. Providing a recommended trading agent to a population: a novel approach. In *Proc. of IJCAI'07*, 2007.

[8] H. Oosterbeek, R. Sloof, and G. van de Kuilen. Cultural differences in ultimatum game experiments: evidence from meta-analysis. *Experimental Economics*, 7:171–188, 2004.

[9] A. Sanfey, J. Rilling, J. Aronson, L. Nystrom, and J. Cohen. The neural basis of economic decision-making in the ultimatum game. *Science*, 300:1755–1758, 2003.